DOCUMENT RESUME

ED 368 793 TM 021 315

AUTHOR Greig, Jeffrey; And Others
TITLE The Development of an Assessment of Scientific Experimentation Proficiency for Connecticut's Statewide Testing Program.
PUB DATE Apr 94
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment; *Experiments; Grade 10; High Schools; High School Students; Pilot Projects; Program Evaluation; Research Methodology; Science Instruction; *Science Tests; Standardized Tests; State Programs; *Test Construction; *Testing Programs; Test Validity
IDENTIFIERS *Connecticut; Connecticut Academic Performance Test; *Performance Based Evaluation

ABSTRACT

Connecticut is developing a new grade 10 statewide testing program known as the Connecticut Academic Performance Test (CAPT). The CAPT reflects the changing nature of standardized assessment by using a variety of assessment formats including performance tasks, open-ended questions, and multiple-choice tests. This paper describes the research that was conducted during the development of the CAPT Science assessment, which includes a performance component. In the design, students are given a hands-on laboratory activity during a 4-week window before the written portion of the test, which then includes some questions on the performance task. The laboratory reports on the performance task are scored only by teachers, and not at the state level. Five performance tasks were developed, and each was completed by nearly 2,000 tenth graders. Approximately 6,000 students responded to follow-up questions without completing the task and thus served as a control group. Results support the feasibility of moving the performance task out of the assessment and into the classroom while maintaining a meaningful link between the two. Four tables and one figure present findings from the pilot studies. Four appendixes contain one task, follow-up questions, and scoring rubrics for each. (Contains 4 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JEFFREY GREIG

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 368 793

The Development of an Assessment of Scientific Experimentation Proficiency for Connecticut's Statewide Testing Program

Jeffrey Greig

Naomi Wise

Michal Lomask

Connecticut State Department of Education

Paper presented at the annual meeting of the American Educational Research Association in New Orleans, Louisiana, April 1994.

TM021315

BEST COPY AVAILABLE

Background

The Connecticut Academic Performance Test (CAPT)

In January 1987 the Connecticut State Board of Education produced a visionary document, called the *Common Core of Learning* (CCL) which describes what Connecticut students should know and be able to do as a result of their education in grades K-12. In an effort to assess the important academic skills and knowledge described in this document, the Connecticut State Department of Education is currently developing a new grade ten statewide testing program known as the Connecticut Academic Performance Test (CAPT). Beginning with the first statewide administration in 1994, the CAPT will annually assess approximately 35,000 students in the content areas of language arts, mathematics, and science as well as on an integrated task. One of the primary purposes of the CAPT is to foster improved instructional practices in the classroom. Results will be used by educators to make informed decisions about their programs and by others to monitor the strength of Connecticut's educational system.

The CAPT reflects the changing nature of standardized testing by using a variety of assessment formats including extended performance tasks, open-ended questions and multiple-choice items. The assessments are set in authentic contexts and require students to integrate their skills and knowledge within and across disciplines.

Issues Related to Science Performance Assessment

The CCL document addresses the issue of scientific experimentation with the following statement, *"At the end of their K-12 education students should be able to identify and design techniques for recognizing and solving problems in science, including the development of hypotheses and the design of experiments to test them - the gathering of data, presenting them in appropriate format, and drawing inferences based upon the results."*

In order to assess students' achievement in the area of scientific experimentation, it was recognized that multiple-choice questions alone are insufficient. In order to assess students' ability to "do" science, they need to be given the opportunity to actually perform a science task. To address this issue, the CAPT Science was designed to include a performance assessment task.

In developing a performance assessment of students' scientific experimentation skills, a number of issues were considered:

1. Consequences

It is known that "what is on the test" shapes what teachers teach and what students learn. This is especially true for high stakes testing programs such as the CAPT. The consequences of the assessment, both intended and unintended, on classroom practice must therefore be of major concern. One

of the main benefits of using performance assessment is that it more closely matches desired instructional and curricular goals. In the case of science instruction, the increased use of hands-on/minds-on activities is an important goal. Therefore, "teaching to the test" is a positive consequence.

2. Generalizability

The stability of students' performance on different tasks is a major concern when using performance assessment. In studying hands-on science assessment activities, Shavelson, Baxter and Pine (1992) have reported high variability in student performance due to task content and format. This is problematic because the number of performance tasks that can be used on a large-scale science assessment is generally very limited due to time and cost constraints.

3. Cognitive complexity

What students are expected to know and be able do in order to perform well on a task, or the cognitive load of the task, is another issue of concern. Previous work (Lomask, Baron and Greig, 1993, Jorgensen, 1993) has shown that science performance tasks of a similar nature to those developed for the CAPT require the application of higher order thinking skills, such as defining problems, designing experiments, and interpreting data. The application of these reasoning skills is highly dependent on familiarity with the content of the task, which might affect the cognitive load of the task.

4. Content quality and coverage

The extent to which an assessment represents the depth and breadth of desired curricular objectives is another concern. Performance tasks have the advantage of assessing a greater depth of understanding of important scientific concepts and thinking skills than objective tests. However, the degree to which a specific science performance task can adequately represent the curriculum is questionable.

5. Cost and efficiency

The cost of performance assessment is higher than the cost of traditional tests. Costs include not only the direct cost of administering and scoring the assessment, but also expenses due to lab equipment and materials. The use of performance assessment in science also presents logistical problems due to the extra time required for administration and the need for proper laboratory facilities and trained administrators. Increased professional development for teachers is also generally required.

The CAPT Assessment of Scientific Experimentation Proficiency

Given the concerns described above, the use of a science performance assessment task on a large-scale assessment presents a considerable challenge. The design of the CAPT assessment of experimentation proficiency attempts to balance these concerns.

In the CAPT design, students are given a hands-on laboratory activity during a four-week window prior to the administration of the written portion of the test. The task presents students with a real-world problem. Students are then instructed to work in pairs to define a specific research question, design and carry out experiments to solve the problem, and draw conclusions based on their results. An example of a CAPT Science performance task is shown in Appendix A.

Students then work individually to write about their experiments and results in the form of a written lab report. These lab reports are not collected and scored at the state level. Rather, teachers are encouraged to score their own students' work and provide them with feedback about their performance. (For the purposes of pilot testing, students' lab reports were collected and scored by the state. Scores were not reported back to students.)

During the written portion of the test, students are given follow-up questions which relate directly to the performance task. These questions are designed to assess students' ability to apply their understanding of scientific experimentation by critiquing sample results. The questions provide students with hypothetical experimental designs, results, and conclusions from the performance task and ask them open-ended questions about the quality of the work shown. An example of the follow-up questions is shown in Appendix B. The results of the follow-up questions are scored at the state level and used in combination with other items to formulate the students' Experimentation score.

This design thus balances the need to encourage the use of hands-on, inquiry-based instruction in the science classroom with technical, logistical and content related concerns. Since students' scores are based on their performance on the follow-up questions (in combination with other experimentation items), rather than on the actual performance task, established equating procedures can be applied to these items. Students' Experimentation scores are then based on their performance on a set of items rather than a single score on one task. This allows for a broader range of experimentation skills to be assessed. This design also gives schools some degree of flexibility in administering the task so that laboratory facilities, equipment and staff are used efficiently thus reducing costs.

This paper describes research that was conducted during the development of the CAPT Science assessment. The following questions related to the validity and generalizability issues described above are addressed:

- Does completing the performance task affect students' scores on the follow-up questions?

- Does the amount of time between completing the performance task and the written test affect students' scores on the follow-up questions?

- What is the relationship between students' scores on their lab reports and on the follow-up questions?

- What is the relationship between students' prior experience in performing labs in their science classes and their scores on the follow-up questions?

- To what degree are the performance tasks comparable?

Method

Subjects

This research was conducted in May 1993 as part of a statewide pilot study for the CAPT program. Approximately 2000 grade ten students from Connecticut public high schools completed one of the five science performance tasks and responded to the corresponding follow-up questions. Approximately 6000 students responded to one of the sets of follow-up questions without completing a performance task, thus serving as a control group.

A stratified random sampling procedure was used to insure state representativeness. Each sample was stratified by Education Reference Groups (ERGs) which is a classification of Connecticut school districts into seven categories based on a variety of socioeconomic factors (e.g., median family income, percentage of high school graduates, percentage of families below poverty). School districts within each ERG classification have generally similar student populations with regard to many of the factors known to affect student achievement. Therefore, the ERG categories are ideal strata for statewide sampling plans.

Materials

Performance Tasks

Five science performance tasks were developed. Each task is a 90-minute laboratory activity that requires students to design and carry out their own experiments to solve a given problem. The problems are set in practical contexts. For instance, if students are examining the factors that affect reaction time, the problem is introduced with a discussion of the time it would take to make a sudden stop to avoid an accident when driving.

Students receive a student booklet containing instructions for the task, space for notes, and several pages to write their lab reports. Students are also provided with all of the materials and equipment needed to complete the task. The instructions describe the problem to be investigated and some suggestions for how to get started. However, students are not provided with step-by-step instructions for carrying-out a particular experiment. Rather, the tasks are designed to allow for a variety of solution paths.

The directions for writing the lab reports guide students to include a clear statement of the problem; a description of the experiment(s); the results (including data presented in charts, tables or graphs); conclusions; and comments about the validity of the conclusions and possible sources of error (see Appendix A). The directions also provide an audience for writing the report. For instance, students might be asked to summarize their results for an article for a consumer magazine.

Follow-up Questions

Accompanying each task are a series of five or six short-answer open-ended questions. These follow-up questions are specific to each task. However, the questions across the five tasks are similar in that they present students with hypothetical experimental designs, results and conclusions from groups of students who supposedly worked on the same task. The follow-up questions then ask students to make judgments about the quality of the work shown (e.g., *Is the problem well defined?, Is this a well designed experiment?, Can any valid conclusions be drawn from the data?*).

The follow-up questions are designed to assess students' ability to apply their understanding of experimentation rather than remember specific details about the experiment they performed earlier. The performance task is intended to provide students with a basis for thinking about the problems presented in the follow-up questions. Students are not allowed to use their lab reports or any notes from the performance task during the follow-up questions.

Procedure

Five science performance tasks and related follow-up questions were piloted. To explore the effect of the amount of time that elapses between completing the performance task and answering the follow-up questions, each task was administered under two conditions: *early* and *late*. In the early condition, students completed the task four weeks prior to the follow-up questions; in the late condition, two weeks prior. In addition, the follow-up questions were given to students who did not complete a performance task at all, to explore whether completing the task was actually related to performance on the follow-up questions. The design of the study and the number of students that took the follow-up questions under each condition are presented in Table 1.

Table 1

Each performance task took 90 minutes to administer. It was required that the tasks be administered in a science laboratory setting, by a certified science teacher. Teachers acted as facilitators, monitoring students to ensure that proper safety procedures were followed and helping to clarify directions as needed. The materials that could be used for each performance task were specified. General lab equipment was provided by the schools; more specific materials required for each experiment were provided by the CSDE and shipped to schools. Students designed and carried out their experiments in pairs. They then wrote about their results individually in the form of a written report.

The follow-up questions were embedded in the written science test that was being piloted. A number of different forms of the science test were piloted; each form had follow-up questions for one of the tasks. The piloting plan was designed so that students who performed a task received the written test form that included the corresponding questions for that task.

Data Collection

Students' reports from the performance tasks were collected to be used as a relatively direct measure of their actual performance on the task. Previous research has shown that lab reports can provide a close approximation to students' actual performance on a lab, as assessed through direct observations (Shavelson, Baxter, Pine, 1992). The reports were scored holistically using a four-point rubric (on a scale from 0 to 3) that takes into consideration such factors as problem definition, appropriateness of experimental design, data collection, validity of conclusions and effective communication (see Appendix C). The rubrics do not penalize students for writing mechanics, such as spelling, punctuation or grammatical errors. The scoring process included the development of scoring rubrics, training and qualifying of scorers, and monitoring of the scoring sessions. Inter-rater agreement was exact for 60% - 79% of the papers, depending on the task; agreement was within one scorepoint for 91% - 100% of the papers, depending on the task. This was judged as an acceptable level of rater agreement.

The follow-up questions were also scored holistically using a four-point rubric (on a scale from 0 to 3) that takes into account the correctness, completeness and appropriateness of students' responses relative to scientific experimentation proficiency. (See Appendix D.)

Observations were made of each of the five tasks being performed in several different classrooms. Students and teachers also responded to questionnaires. The following questions

from the student questionnaire were considered as part of this study: *Have you performed lab experiments in your science classes before?; If you have performed lab experiments in science class before, were you ever asked to design your own experiment?*

Findings

Does completing the performance task affect students' scores on the follow-up questions?

This question addresses the validity of the link between the performance tasks and follow-up questions. To justify the time and cost of having high schools administer a performance task to all of their tenth grade students prior to the written assessment, it must be shown that the follow-up questions are meaningfully linked to the performance task. As predicted, our pilot data shows that students who completed a performance task scored significantly higher on the corresponding follow-up questions for each of the five tasks (Table 2). Individual t-tests were run; the Dunn (1961) procedure was then applied to adjust the overall alpha level for the number of comparisons being made (.05 divided by 5 comparisons, resulting in an adjusted p-value of .01). The consistent significant effects of performing the lab shown in Table 2 support a meaningful link between the performance task and the follow-up questions.

Table 2

Does the amount of time between completing the performance task and the written test affect students' scores on the follow-up questions?

When the assessment is administered statewide, schools will be allowed considerable flexibility in scheduling the performance task according to the availability of labs and instructors. One question that arises is whether the amount of time that elapses between performing the task and the written test affects students' scores on the follow-up questions. Table 3 compares scores on the follow-up questions for students who performed the task four weeks prior to the written test versus two weeks prior.

Table 3

The findings are mixed and somewhat surprising. For two of the five tasks, there were significant differences between the two groups. Counter to our expectations, students who completed the task four weeks prior to the test scored significantly higher on the follow-up questions than students who completed the task two weeks prior. For the other three tasks, there were no significant differences between the two conditions. These mixed findings negate concerns that allowing a large amount of time between the performance task and answering the follow-up questions will detract from student performance.

One might have predicted that the less time between the performance tasks and the follow-up questions, the better student performance on the questions. However, as described above, the follow-up questions do not require students to remember specific details about the experiment they performed earlier; rather the task is intended to provide a basis for critiquing others' approaches to the problem. The fact that students' scores on the follow-up questions were significantly higher under the four week condition is worthy of further study.

What is the relationship between students' scores on their lab reports and their scores on the follow-up questions?

This is another question that relates to the validity of the link between the performance tasks and follow-up questions. For the purposes of this study, the lab reports were treated as relatively direct measures of students' performance on the actual tasks. Correlations between students' scores on the lab reports and their scores on the follow-up questions range from .27 to .44 , depending on the task (.42, .27, .33, 44 and .34 for Tasks 1 through 5, respectively). These are only moderate correlations, which is not surprising since the questions are intended to measure somewhat different skills than the performance tasks. In the performance tasks, students are asked to design their own experiments, carry them out and report their results. The follow-up questions require students to reflect on the work of others by critically evaluating sample experimental designs, results and conclusions. These are different but related skills, both of which were judged to be important to promote and assess.

What is the relationship between the extent of students' prior experience performing labs in their science classes and their scores on the follow-up questions?

To further explore the validity of using the follow-up questions to assess experimentation proficiency, performance on the follow-up questions was compared for students who reported through a student questionnaire having performed labs in their science classes *many times before* versus *just a few times*. Table 4 shows that students performing labs frequently in their science classes scored significantly higher on the follow-up questions. There is a significant effect for each of the five tasks.

Table 4

Students were also asked to report on the frequency with which they are required to design their own experiments in their science classes. Scores on the follow-up questions were compared for students who reported having been asked to design their own science lab experiment *many times* versus *only a few times*. There was a significant effect for Task #1--contrary to expectations, students who reported "a few times" scored significantly higher on the follow-up questions than those who reported "many times" ($t=3.24$, $p<.01$). For the other four tasks, there were no significant effects. Thus, there is no evidence of the expected relationship between prior experience in designing experiments and performance on the follow-up questions. These findings suggest that more research is needed to further explore the nature of what is being assessed through the follow-up questions.

Are the performance tasks comparable?

One of the greatest challenges in developing alternative methods for assessing science experimentation proficiency is creating tasks of equal difficulty to be used from year to year, so that growth in student achievement can be measured.

Figure 1 compares the frequency distributions of scores on the lab reports for each of the five tasks piloted. The lab reports were scored on a 4- point scale (0 to 3). Figure 1 shows that the distributions of scores across tasks are similar, but the tasks are not comparable enough to be used for reporting changes in student achievement in science experimentation proficiency from year to year. This finding is not surprising since existing research (referenced earlier in this

paper) has found high variability in student performance among different hands-on science tasks.

Figure 1

Pilot data indicates that there are some differences in the difficulty levels of various follow-up questions developed for each of the tasks. The individual items range in difficulty from .68 to 1.55 on a scale from 0 to 3. However, unlike a single performance task, sets of items can be equated using established procedures.

In addition to the follow-up questions related to the performance tasks, several other experimentation questions unrelated to the task will appear on the science test each year. There will be a total of eight experimentation items on each form of the test and those eight items will comprise students' experimentation score. All items contributing to the Experimentation score will be calibrated using an IRT one-parameter partial credit model. They will then be reported as a scale score. This provides the opportunity to make adjustments for differences.in the difficulty of underlying items so that the scale scores that are reported are comparable from year to year.

Discussion

The use of performance assessment on a large-scale testing program requires trade-offs. Connecticut's new grade ten assessment will test approximately 35,000 students each year. The CAPT assessment of scientific experimentation proficiency is designed to promote hands-on inquiry-based science instruction while still addressing a variety of technical and logistical assessment concerns. The solution chosen was to move the performance task out of the assessment and into the classroom, while maintaining a meaningful link between the two.

The use of the follow-up questions does not allow for the measurement of students' ability to design and carry out their own scientific experiments. Rather, the ability of students to apply scientific thinking to critically evaluate the work of others is assessed. These are both important reasoning skills. Given the constraints of a large-scale assessment, a decision was made to assess the latter while promoting the use of the former in instruction. More specifically:

• Through the CAPT design, the use of hands-on inquiry-based laboratory activities in the science classroom is encouraged. Student performance on the follow-up questions, used as basis for students' Experimentation scores, is meaningfully linked to completing the performance task. In addition, student performance is correlated with consistent use of laboratory activities in the classroom. However, the extent to which this design will actually affect instruction is still unknown and is in need of further study.

• As expected, findings from this study confirmed previous research that has found considerable variability among science performance tasks. The CAPT design then takes this variability into consideration. Since scores on the lab reports from the tasks are not being reported, the same psychometric standards for equivalence need not be applied. The follow-up questions can be equated from year to year using established equating procedures.

• The use of an inquiry-based laboratory activity along with follow-up questions (in combination with other experimentation items) in the assessment design allows for a broader coverage of important curricular objectives. Using the results of a single performance task as the sole basis for students' Experimentation scores could result in a narrowing of the curriculum.

• By placing the performance task outside of the assessment, the need to standardize administration procedures is reduced. Logistical difficulties and costs related to scoring and laboratory equipment are therefore also reduced.

References

Dunn, O.J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association*, 56, 52-64.

Jorgensen, M., (1993). Assessing Habits of Mind: Performance-Based Assessment in Science and Mathematics. Columbus, Ohio: ERIC Clearinghouse for Science, Mathematics and Environment Education at the Ohio State University.

Lomask, M., Baron, J., and Greig J., (1993). Alternative Assessment in Science, in J. Baron (Ed.) Assessment as an Opportunity to Learn, Connecticut State Department of Education.

Shavelson R. J., Baxter G. P., and Pine J., (1992). Performance Assessment - Political Rhetoric and Measurement Reality. *Educational Researcher*, May 1992, 22-27.

Table 1

Number of Students Completing the Follow-Up Questions Under Different Conditions

	Number of Ss		
	Performance Task		No Performance Task
	Early	Late	
Task #1	214	226	1274
Task #2	196	181	1173
Task #3	184	206	1016
Task #4	215	176	1109
Task #5	210	226	1051

Table 2

Mean Total Score on Follow-up Questions for Students Completing and Not Completing a Performance Task

	Performance Task	No Performance Task	
	Mean (SD)	Mean (SD)	t
Task #1[a]	4.07 (2.82)	3.14 (2.55)	6.14*
Task #2[a]	5.36 (3.33)	4.82 (3.66)	2.66*
Task #3[a]	5.55 (3.37)	4.86 (3.67)	3.35*
Task #4[b]	4.94 (3.36)	4.33 (3.75)	2.97*
Task #5[a]	5.62 (3.44)	4.35 (3.57)	6.41*

* Significant at .05 level after Dunn adjustment for multiple comparisons

a Maximum score = 15

b Maximum score = 18

Table 3

Mean Total Score on Follow-Up Questions for Students Completing a Performance Task Two Weeks Versus Four Weeks Earlier

	Amount of Time Between Performance Task and Follow-Up Questions		t
	Two Weeks M (SD)	Four Weeks M (SD)	
Task #1 [a]	3.72 (2.56)	4.44 (3.04)	2.69*
Task #2 [a]	5.34 (3.49)	5.37 (3.19)	0.09
Task #3 [a]	5.75 (3.23)	5.32 (3.53)	1.26
Task #4 [b]	5.14 (3.67)	4.77 (3.08)	1.06
Task#5 [a]	4.90 (3.04)	6.40 (3.68)	4.61*

* Significant at .05 level after Dunn adjustment for multiple comparisons

[a] Maximum score = 15

[b] Maximum score = 18

Table 4

Mean Total Score on Follow-Up Questions for Students Who Reported Having Done Labs in Their Science Classes Many Times Versus Just a Few Times Before

	Performed Labs in Your Science Classes Before?		
	Many Times	Just a Few Times	
	M (SD)	M (SD)	t
Task #1 [a]	4.62 (2.88)	3.90 (2.58)	2.48*
Task #2 [a]	5.94 (3.46)	4.94 (2.89)	2.77*
Task #3 [a]	6.53 (3.45)	5.19 (2.81)	3.86*
Task #4 [b]	5.60 (3.65)	4.35 (2.85)	3.56*
Task #5 [a]	6.56 (3.44)	4.96 (3.06)	4.82*

* Significant at .05 level after Dunn adjustment for multiple comparisons

[a] Maximum score = 15

[b] Maximum score = 18

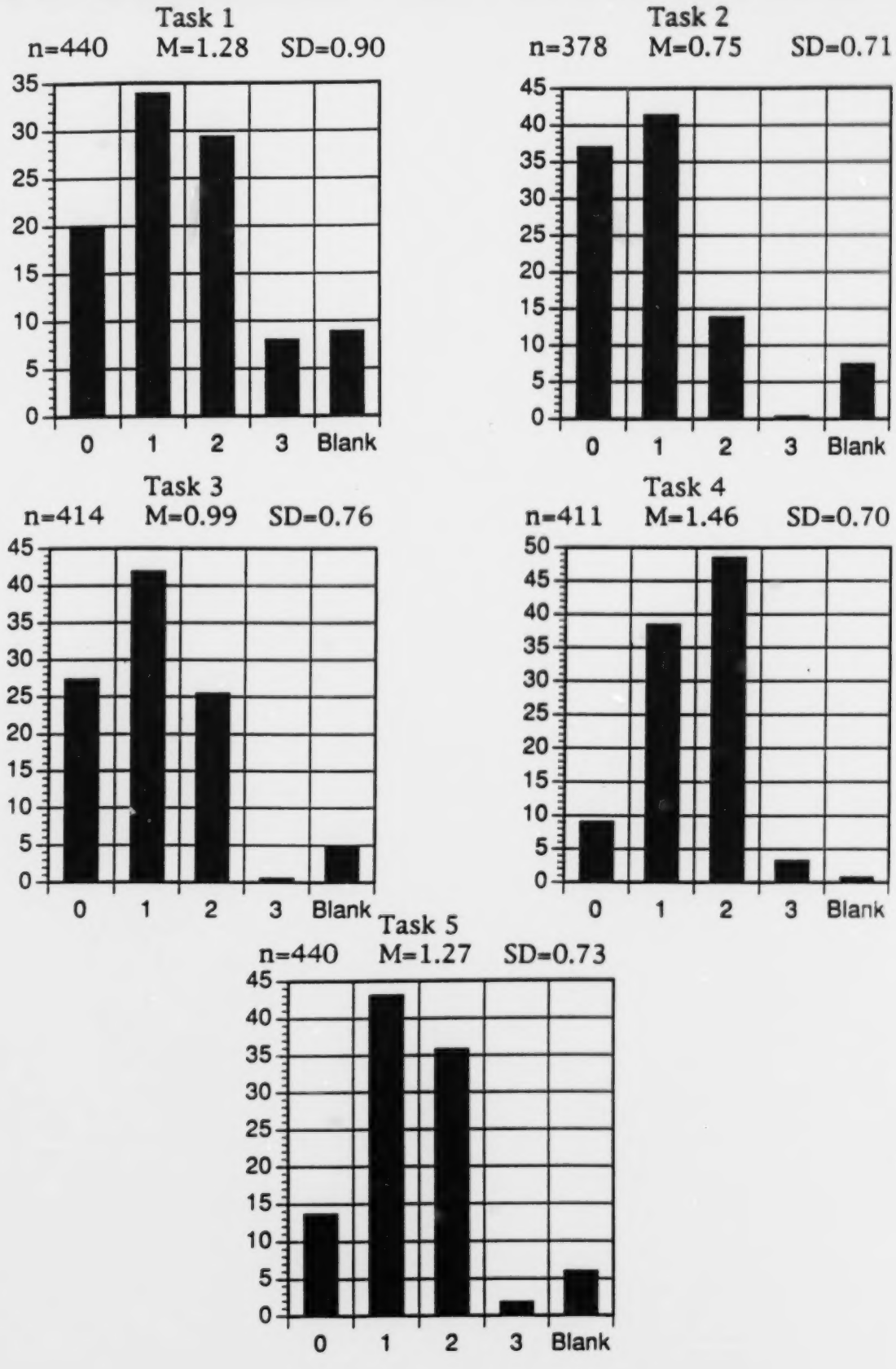


Figure 1 Frequency distribution of scores on the lab reports.

Appendix A

Sample CAPT Science Performance Task

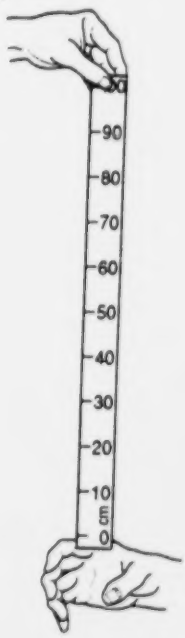
QUICK REFLEXES

Suppose you are driving on a highway at 55 miles per hour. Suddenly you see a crate fall off of a truck in front of you. How long will it take you to move your foot from the gas pedal to the brake pedal? The time that it takes you to react is called your *reaction time.* What factors affect your reaction time?

In this activity you will be exploring the reaction time of you and your partner(s) and various factors that might affect human reaction time.

Your Task

Today you and your partner will conduct an experiment to determine your reaction times and to find ways to improve your reaction times. You will be provided with a meter stick, a chart for determining reaction times (see the next page) and the directions in this booklet.



Steps to Follow

1. An easy way to measure reaction time is with a meter stick. Hold the meter stick by the upper end, in a vertical position, with the zero sign at the bottom. Ask the student whose reaction time is being tested to place his/her thumb and forefinger at the lower end of the stick, without touching it. Now drop the stick and let the student catch the stick. The position of the student's fingers on the meter stick marks the distance the stick fell, and it can serve as a basis for estimating (or calculating) the reaction time of that student.

2. Determine your partner's reaction time, and have your partner determine your reaction time. First have your partner catch the meter stick. Record the number of centimeters at the spot where the stick was caught. Then use the chart on the next page to determine your partner's reaction time. Now have your partner determine your reaction time.

3. State one or more specific problems or questions related to reaction time (e.g. How might reaction time be improved? Is there a difference between reaction time of your right and left hand?).

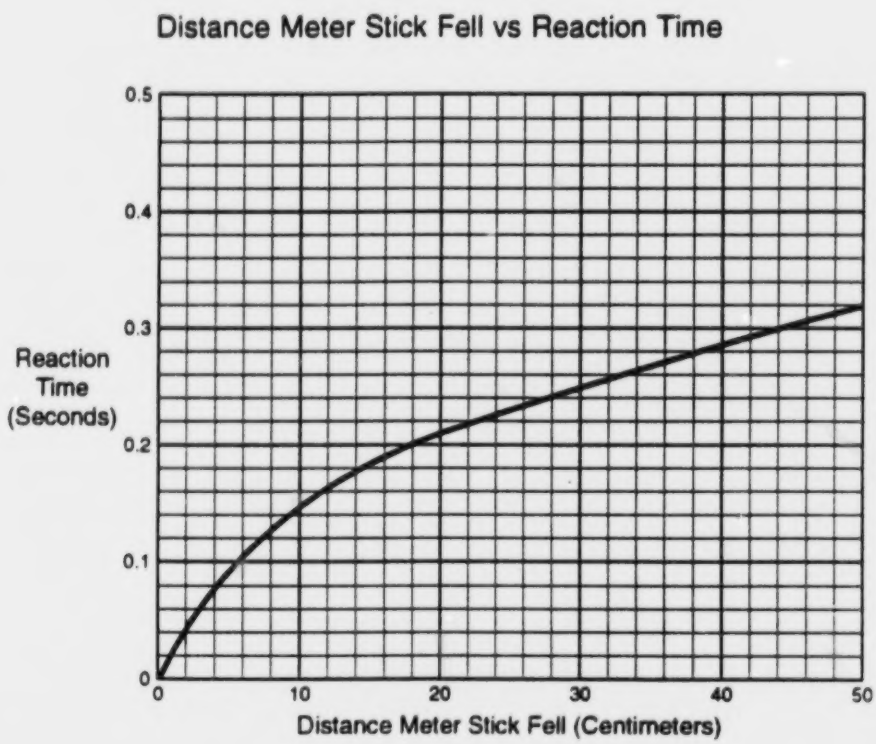
4. Design an experiment to solve the problem(s). Write your experimental design on the space provided. Show your experimental design to your teacher before you begin your experiment.

QUICK REFLEXES

5. After receiving approval from your teacher, work with your partner(s) to carry out your experiment. Your teacher's approval does not necessarily mean that your teacher thinks your idea is good. It simply means that in your teacher's judgment your experiment is not dangerous or likely to cause an unnecessary mess.

6. While experimenting, take notes on the attached pages. Space is also provided for charts, tables or graphs. Your notes will not be scored, but they will be helpful to you later as you work independently to write about your experiments and results. You must keep your own lab notes because you will not work with your partner(s) when you write your article.

7. As time permits, design and carry out experiments to solve other reaction time problems. Be sure to get your teacher's approval before conducting an experiment.



Sample CAPT Science Performance Task: Written Report

Directions for Writing Your Report

You will now summarize your experiments and results in the form of an article for an automobile-safety magazine. You may use the lab notes you took while working with your partner(s). You may wish to write a first draft of your report on scratch paper, but your final copy should be written on the following pages in this booklet. Space for charts, tables or graphs is provided.

Your article should include:

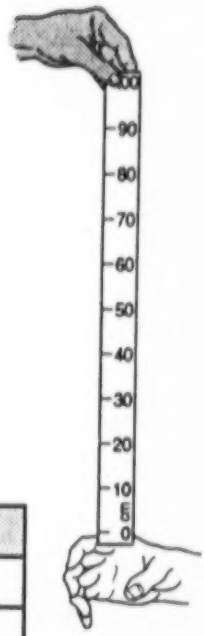
- a clear statement of the problem(s) you investigated;
- a description of the experiment(s) you carried out;
- the results of your experiments (including data presented in the form of charts, tables or graphs);
- your conclusions from the experiments; and
- comments about how valid you think your conclusions are. (In other words, how much confidence do you have that your results are accurate? What errors may have affected your results?)

Appendix B

Sample CAPT Follow-up Questions

QUICK REFLEXES

A group of students measured each other's reaction time, using only a meter stick.



Shantay and Al were lab partners. Shantay held the tip of a meter stick between her thumb and forefinger, and Al held his hand ready to grasp the other end of the meter stick.

When Shantay let go of the meter stick, Al tried to catch it before it hit the floor. They recorded the distance it dropped. Then Al held the meter stick for Shantay.

They repeated the experiment four times. The table shows their results.

	Trial 1	Trial 2	Trial 3	Trial 4
Al	28 cm	34 cm	29 cm	36 cm
Shantay	25 cm	20 cm	22 cm	28 cm

1. When Shantay wrote her lab report, she concluded that their experiment proves that girls have quicker reaction times than boys. Do you think this is a valid conclusion? Explain why or why not.

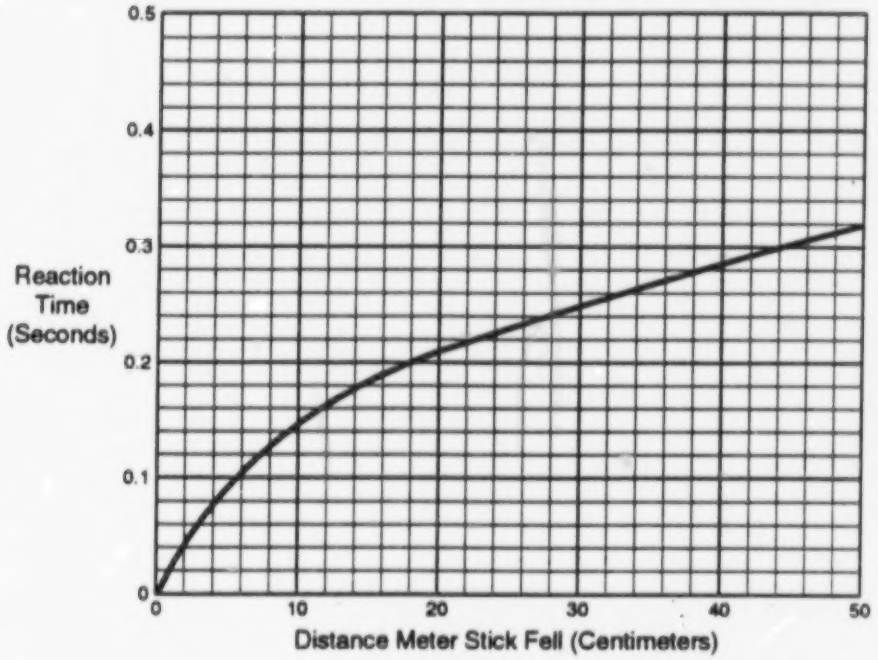
2. What other conclusions can you draw from this data?

3. Cathy and Ben, two other students in the class, studied a different aspect of reaction time. The table below shows their results.

	Right Hand	Left hand
Cathy	23 cm	20 cm
Ben	25 cm	26 cm

Cathy and Ben concluded that left-handed people have faster reaction time than right-handed people. Do you think this is a valid conclusion? Explain why or why not.

QUICK REFLEXES



4. A student named Juanita was also experimenting with meter sticks, and she obtained the following results:

	Trial 1	Trial 2	Trial 3	Trial 4
Juanita	24 cm	30 cm	20 cm	22 cm

Using this information and the graph above, determine Juanita's reaction time in seconds. Explain how you determined your answer.

5. Suppose your class is experimenting with meter sticks. State a problem of your own related to reaction time, and design a complete experiment to answer the question.

Appendix C

General Scoring Rubric for CAPT Science Performance Task Lab Reports

Excellent Performance

The response reflects excellent problem-solving and science process skills. The problem is defined clearly. An appropriate experimental design has been selected and employed rigorously. Reasoning is logical and explained thoroughly. Inferences and conclusions are supported by appropriate observations. There are few if any misconceptions or errors, and none of them are serious. The methods and results are communicated clearly enough that a reader could easily repeat the experiment.

Proficient Performance

The response reflects proficient problem-solving and science process skills. The problem is defined adequately. An experimental design is evident although it may not be completely appropriate and/or may not be employed rigorously. Reasoning is generally logical. Most inferences and conclusions are supported by observations. There are few if any serious misconceptions as well as other errors. The methods and results are communicated clearly enough for a reader to understand what the student has done, but there may be omissions and/or inconsistencies that would hinder a reader from being able to repeat the experiment easily.

Marginal Performance

The response reflects marginal problem-solving and science process skills. The problem may be defined poorly. There may be some evidence of an experimental design, but it may be inappropriate and/or may not have been employed well. Reasoning may contain significant flaws. There may be some inferences and conclusions that can be supported by observations, but others may not be supportable, and those that are supportable may not have been supported adequately. There may be some evidence of serious misconceptions as well as other errors. An attempt has been made to communicate the student's methods and results, but a reader would probably have significant difficulty repeating the experiment.

Unsatisfactory Performance

The response reflects unsatisfactory problem-solving and science process skills. The definition of the problem may be very limited or altogether missing. There is little if any evidence of an experimental design. Reasoning may be illogical, or it may contain numerous errors. There may be few if any inferences or conclusions, and those that appear may not be supportable. There may be numerous and serious misconceptions as well as other errors. There may be little evidence that the student tried to communicate his or her methods and results. Any attempt that has been made to communicate the methods and results would not enable a reader to reproduce the experiment.

Appendix D

General Scoring Rubric for Open-ended Follow-up Questions

Excellent Performance

The response is an excellent answer to the question. It is correct, complete, appropriate and contains elaboration and/or evidence of higher-order thinking and relevant prior knowledge. There is no evidence of misconceptions. Minor factual errors will not necessarily lower the score.

Proficient Performance

The response is a proficient answer to the question. It is generally correct, complete and appropriate although minor inaccuracies may appear. There may be limited evidence of elaboration, extension, higher-order thinking and relevant prior knowledge, or there may be significant evidence of these traits but other flaws (e.g. inaccuracies, omissions, inappropriateness) may be more than minor. There may be evidence of minor misconceptions.

Marginal Performance

The response is a marginal answer to the question. While it may contain some elements of a proficient response, it is inaccurate, incomplete and/or inappropriate. There is little evidence of elaboration, extension, higher-order thinking or relevant prior knowledge. There may be evidence of significant misconceptions.

Unsatisfactory Performance

The response, although on topic, is an unsatisfactory answer to the question. It may fail to address the question, or it may address the question in a very limited way. There may be no evidence of elaboration, extension, higher-order thinking or relevant prior knowledge. There may be some evidence of serious misconceptions.